1. **Model Building and Interpretation**

**Understanding Business/Social Opportunity:**

During the study of causes/factors responsible for heart attack, I believe that it is a social opportunity to inform the people about the health concerns and the major disadvantages related to smoking cigarettes, high cholesterol level, high blood pressure etc. meanwhile, arranging awareness program informing the benefits of regular exercises and good food habits.

It is also the business opportunity for the pharmaceutical industries and health instructors/ dieticians to create wealth by manufacturing drugs and suggesting proper diet and exercises to their customers.

**Exploratory Data Analysis:**

* Age is a continuous variable with minimum value of 32 and maximum value of 70. It has mean of 49.58.
* The data set contains the “NA” values for the columns “educations”, “cigsPerDay”, “BPMeds”, “totChol”, “BMI”, “heartRate”, “glucose”.
* totChol has a minimum value of 107 and maximum value of 696. It also has a mean value of 236.7.
* glucose has a minimum value of 40 and a maximum value of 394. The mean value of glucose is 81.96.
* heartRate is also found to be a very important variable as it has large variation between minimum and maximum value. The minimum value is 44 and maximum value is 143. The mean value of heart rate is 75.88.
* diaBP is also a variable of consideration. It has a minimum value of 48 and maximum value of 142.5. The mean value is 82.9.

**Univariate Analysis:**

* Variable age has a right skewed distribution.
* Variable BMI also has right skewness.
* Variable cigsPerDay has highly right skewed distribution.
* Variable diaBP has normal distribution.
* Variable glucose is also highly right skewed.
* Variable heartRate also has right skewed distribution.
* Variable sysBP has right skewness.
* Variable totChol is also right skewed.

**Missing Value Treatment:**

* Variable heartRate has 0.02% of missing values.
* Variable BMI has 0.45% of missing values.
* Variable cigsPerDay contains 0.68% of missing value.
* Variable totChol contains 1.18% of missing values.
* Variable BPMeds contains 1.25% of missing values.
* Variable education contains 2.48% of missing values.
* Variable glucose contains 9.15% of missing values.
* In order to have a good model building, it is required to treat these missing values.
* For treatment of missing values, imputation method is deployed.
* Since the variable education does not contain any outlier, therefore for its missing value treatment, imputation by mean is done.
* Rest all the other variables has outliers therefore to treat them imputation by median is done.
* After imputing the missing values, again the plot of missing values is done to verify whether or not any missing value exist in the data set.

**Outlier Treatment:**

* The boxplot of each variable shows that the variables BMI, cigsPerDay, diaBP, glucose, heartRate, sysBP and totChol contains the outlier.
* The treatment of outliers involves replacing the outliers by capped values meaning replacing those values which lie outside the lower limit by 5th percentile and those which lie above the upper limit by the value of 95th percentile.
* After outlier treatment, boxplot of all the variables BMI, cigsPerDay, diaBP, glucose, heartRate, sysBP and totChol is plotted again to verify whether the outliers exist in the data set or not.

**Build Various Models:**

Various models were built on the data set provided by the medical practitioners. The purpose behind building different models is to test the accuracy of predicting risk of getting heart attack. The various models that were built are as follows:

1. **Simple Logistic Regression Model:**

* Missing value and outlier treatment was done.
* Variable conversion is done to obtain the variables in proper format.
* In order to build the logistic regression model, library “lmtest” is utilized.
* Under “lmtest”, “glm” function is used to build the model.
* Model designing is done on all the independent variables (excluding TenYearCHD, which is our dependent variable).
* Summary of the model shows that the variables male, age, cigsPerDay, prevalentStroke, sysBP and glucose are important variables as there p value less than 5%.
* Again the model is designed considering the important variables and accuracy is calculated from confusion matrix (Table-1).
* Confusion matrix is created by testing the model on test data set and considering 0.5 as threshold i.e. above 0.5 probability it will predict output as 1 and below it the output is predicted as 0.
* ROC curve is shown in Fig.1.
* From Table-1, it is concluded that the model is able correctly predict 2525 observation out of 2968 observations.
* The various performance measures are as follows:
  + Accuracy = 85.07%
  + Specificity = 58%
  + Sensitivity = 85.53%
  + AUC = 73.32%

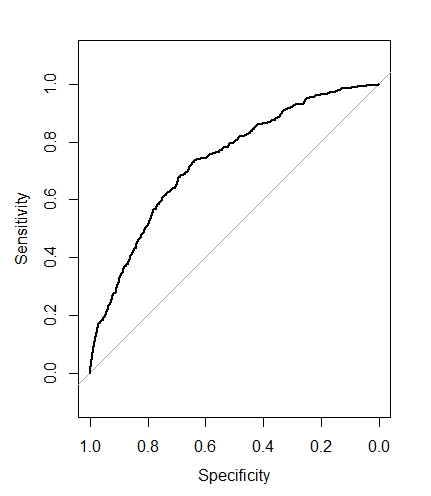


Fig.1: ROC curve

Table-1: Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 2496 | 21 |
| 1 | 422 | 29 |

1. **Logistic Regression with SMOTE Model:**

* Since the data set given to us has an output variable TenYearCHD which is a categorical variable containing 0’s and 1’s.
* Missing value treatment and outlier treatment is done.
* Variable conversion is done to obtain the variables in proper format.
* Analyzing the data set, it was found that the data set was a little unbalanced i.e. containing 85:15 distribution of 0’s and 1’s respectively.
* Therefore applying SMOTE analysis to obtain a distribution of 70:30.
* The SMOTE data is then divided into train and test data set to carry out further model building exercise.
* In order to build the logistic regression model with SMOTE, library “lmtest” is utilized.
* Under “lmtest”, “glm” function is used to build the model.
* Model designing is done on all the independent variables (excluding TenYearCHD, which is our dependent variable).
* Confusion matrix is created by testing the model on test data set and considering 0.5 as threshold i.e. above 0.5 probability it will predict output as 1 and below it the output is predicted as 0.
* ROC curve is shown in Fig.2.
* From Table-2, it can be observed that the model is able to correctly predict 83% of observations.
* The various performance measures are as follows:
  + Accuracy = 82.79%
  + Specificity = 80.34%
  + Sensitivity = 83.40%
  + AUC = 86.55%

Table-2: Confusion Matrix for LR with SMOTE

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 31947 | 1863 |
| 1 | 6358 | 7617 |

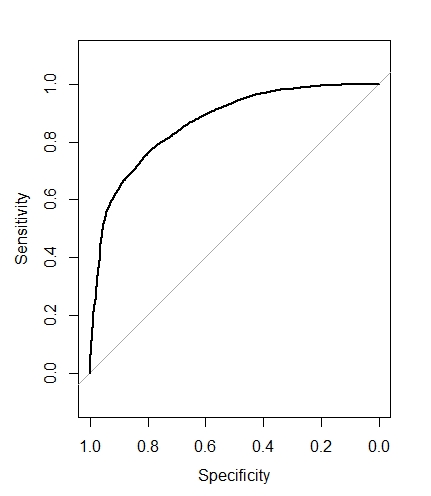


Fig.2: ROC curve for LR with SMOTE

1. **Logistic Regression with SMOTE and cross validation Model:**

* On the above build logistic regression model using SMOTE, the K fold cross validation is done.
* K fold cross validation is done by assuming K = 10.
* Confusion matrix is created by testing the model on test data set and considering 0.5 as threshold i.e. above 0.5 probability it will predict output as 1 and below it the output is predicted as 0 (Table-3).
* ROC curve is shown in Fig.3.
* From Table-3, it is found that the model is able to correctly predict 83% of the observations.
* The various performance measures are as follows:
  + Accuracy = 83.17%
  + Specificity = 55.43%
  + Sensitivity = 94.64%
  + AUC = 86.55%

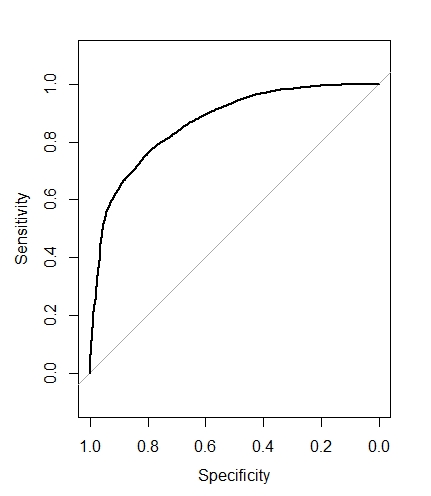


Fig.3: ROC curve for LR with SMOTE and cross validation

Table-3: Confusion Matrix for LR with SMOTE and cross validation

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 13713 | 2669 |
| 1 | 777 | 3320 |

1. **Simple Naïve Bayes Model:**

* First of all the missing value treatment and outlier treatment was done on the data set.
* The data set is divided into train and test data set.
* To build the Naïve Bayes model, library “e1071” was utilized.
* To build the Naïve Bayes model, the output variable is converted into factor.
* The model was build considering all the independent variables.
* Confusion matrix is calculated/obtained (Table-4).
* ROC curve is shown in Fig.4.
* Various performance measures are calculated as follows:
  + Accuracy = 81.68%
  + Specificity = 27.46%
  + Sensitivity = 91.38%
  + AUC = 71.68%
* From Table-4, it is seen that the model is able to correctly predict 1039 observation out of 1272.

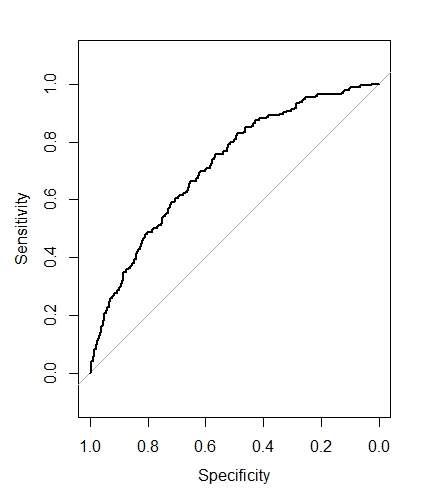
****

Fig.4: ROC for NB

Table-4: Confusion Matrix for NB

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 986 | 140 |
| 1 | 93 | 53 |

1. **Naïve Bayes with SMOTE Model:**

* First of all missing value treatment and outlier treatment was done on the data set.
* Data is analyzed for ratio of data distribution. It was found that the data set contains 85% of 0’s and 15% of 1’s
* SMOTE technique is applied to convert the above data distribution in a 70:30.
* The data set is divided into train and test data set.
* To build the Naïve Bayes model, library “e1071” was utilized.
* To build the Naïve Bayes model, the output variable is converted into factor.
* The model was build considering all the independent variables.
* Confusion matrix is calculated/obtained (Table-5)
* ROC curve is shown in Fig.5.
* From Table-5, it is found that the accuracy of the model is 79.63% i.e. the model can predict approximately 80% times the correctly.
* Various performance measures are calculated as follows:
  + Accuracy = 79.63%
  + Specificity = 59.13%
  + Sensitivity = 88.10%
  + AUC = 86.32%

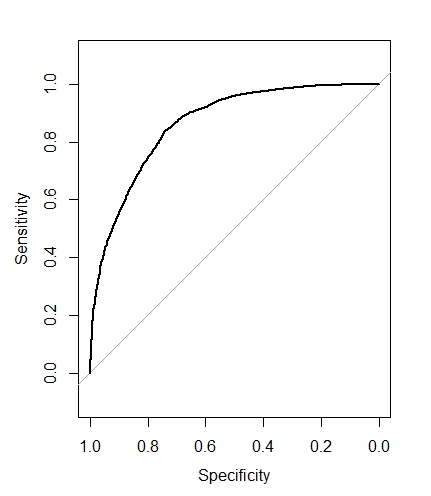


Fig.5: ROC for NB with SMOTE

Table-5: Confusion Matrix for NB with SMOTE

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 12766 | 2448 |
| 1 | 1724 | 3541 |

1. **Naïve Bayes with SMOTE and cross validation Model:**

* First of all missing value treatment and outlier treatment was done on the data set.
* Data is analyzed for ratio of data distribution. It was found that the data set contains 85% of 0’s and 15% of 1’s
* SMOTE technique is applied to convert the above data distribution in a 70:30.
* The data set is divided into train and test data set.
* To build the Naïve Bayes model, library “e1071” was utilized.
* On the SMOTE model K fold cross validation is done.
* K fold cross validation is done by assuming K = 10.
* The model was build considering all the independent variables.
* Confusion matrix is calculated/obtained (Table-6)
* ROC curve is shown in Fig.6.
* From Tabel-6, it is seen that the model has a very high sensitivity of 94.40% meaning that the model is highly sensitive in predicting the true positive rate.
* Various performance measures are calculated as follows:
  + Accuracy = 83.84%
  + Specificity = 58.29%
  + Sensitivity = 94.40%
  + AUC = 87.12%

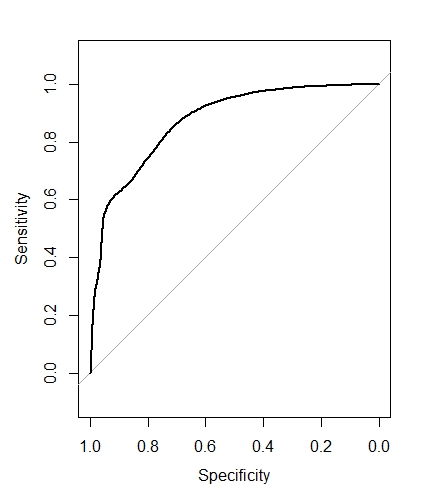


Fig.6: ROC for NB with SMOTE and cross validation

Table-6: Confusion Matrix for NB with SMOTE and cross validation

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 13678 | 2498 |
| 1 | 812 | 3491 |

**Interpretation of Best Model:**

Following models were built based upon the data set to predict the heart attack rate. Outcomes of all the performance measures for different model is listed in Table-7

Table-7: Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Model Name** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| 1 | Simple Logistic Regression | 85.07% | 85.53% | 58% | 73.32% |
| 2 | Logistic Regression with SMOTE | 82.79% | 83.40% | 80.34% | 86.55% |
| 3 | Logistic Regression with SMOTE and Cross Validation | 83.17% | 94.64% | 55.43% | 86.55% |
| 4 | Simple Naïve Bayes | 81.68% | 91.38% | 27.46% | 71.68% |
| 5 | Naïve Bayes with SMOTE | 79.63% | 88.10% | 59.13% | 86.32% |
| 6 | Naïve Bayes with SMOTE and Cross Validation | 83.84% | 94.40% | 58.29% | 87.12% |

* In order to interpret the best model, first priority is given to AUC, second priority is given to sensitivity.
* Therefore, according to Table-7, our first best model is Naïve Bayes with SMOTE and cross validation.
* Second best model based on sensitivity is logistic regression with SMOTE and cross validation.

1. **Model Tuning**
2. **Simple Random Forest Model:**

* Data set is first treated for missing value and outliers.
* Variable conversion is done to transform the variables into correct dimension.
* Data set is then split into train and test data.
* Random forest is built on train data set and its performance is tested on test data set.
* Confusion matrix is shown in Table-8.
* Random forest error v/s trees is shown in Fig.7.
* Minimum number of trees for tuning is found to be approximately 25 (Fig.7).
* Important variables are found to be age, education, cigsPerDay, totChol, sysBP, diaBP, BMI, heartRate, glucose.
* Minimum OOB error is found to 0.03%
* From Table-8, it is found that the model sensitivity is 85.53% in predicting the true positive rate.

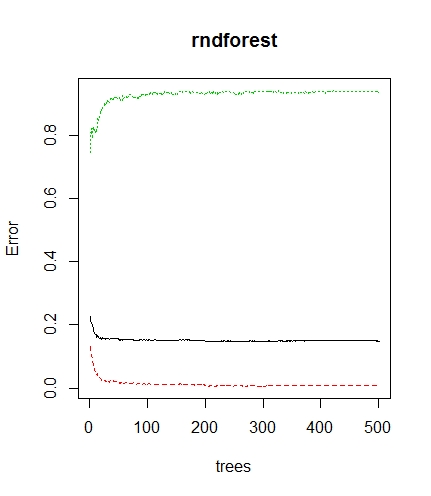
****

Fig.7: Simple Random Forest error v/s trees

Table-8: Confusion Matrix on testDS for simple random forest

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1064 | 15 |
| 1 | 180 | 13 |

* Performance measures on test data set are as follows (Table-8):
  + Accuracy = 84.67%
  + Sensitivity = 85.53%
  + Specificity = 46.42%
  + AUC = 70.45%

**Random Forest with SMOTE Model:**

* First of all missing value treatment and outlier treatment was done on the data set.
* Data is analyzed for ratio of data distribution. It was found that the data set contains 85% of 0’s and 15% of 1’s
* SMOTE technique is applied to convert the above data distribution in a 70:30.
* The data set is divided into train and test data set.
* Random forest is built on train data set and its performance is tested on test data set.
* Confusion matrix is shown in Table-9.
* Random forest with SMOTE, error v/s trees is shown in Fig.8.
* The sensitivity of the model is found to be almost 98% i.e. model is capable of predicting 98% of times the true positive rate.
* Performance measures on test data set are as follows (Table-9):
  + Accuracy = 98.34%
  + Sensitivity = 97.71%
  + Specificity = 100%
  + AUC = 99.95%

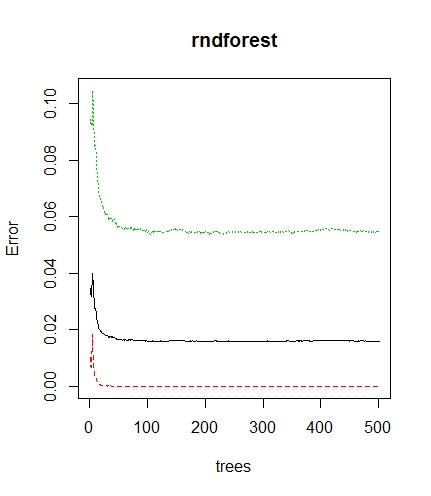
****

Fig.8: Random Forest with SMOTE error v/s trees

Table-9: Confusion Matrix on testDS for SMOTE random forest

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 14490 | 0 |
| 1 | 340 | 5649 |

**Interpretation of Best Model:**

* Various models were built on the given data set.
* Performance measures for all the model is tabulated as shown in Table-10.
* To identify the best model, first priority is given to the AUC i.e. the model with highest AUC value is considered as the best model.
* Second priority is given to the sensitivity i.e. the model with highest value of sensitivity is considered as the second best model.

Table-10: Final Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No.** | **Model Name** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| 1 | Simple Logistic Regression | 85.07% | 85.53% | 58% | 73.32% |
| 2 | Logistic Regression with SMOTE | 82.79% | 83.40% | 80.34% | 86.55% |
| 3 | Logistic Regression with SMOTE and Cross Validation | 83.17% | 94.64% | 55.43% | 86.55% |
| 4 | Simple Naïve Bayes | 81.68% | 91.38% | 27.46% | 71.68% |
| 5 | Naïve Bayes with SMOTE | 79.63% | 88.10% | 59.13% | 86.32% |
| 6 | Naïve Bayes with SMOTE and Cross Validation | 83.84% | 94.40% | 58.29% | 87.12% |
| 7 | Simple Random Forest | 84.67% | 85.53% | 46.42% | 70.45% |
| 8 | Random Forest with SMOTE | 98.34% | 97.71% | 100% | 99.95% |

**Conclusion from Model Building:**

**The best model is Random Forest with SMOTE model as it has the highest value of AUC (Area Under Curve) i.e. 99.95%.**

**Recommendation:**

* Patients with high level of glucose and cholesterol are advised to do exercise regularly.
* Such patients are also advised to maintain the level of sugar by following a proper diet chart prescribed by the dietician.
* Patients who are frequent smoker and diabetic, they are advised to walk daily atleast 5km.
* Patients who are having medical history are advised to take the medicines properly on time and consult with the doctors about their health.